統計學

Fall 2025

政治大學統計系余清祥

2025年9月16日

第三章：敘述統計(統計數值)

http://csyue.nccu.edu.tw

Chapter Contents

CENGAGE

Numerical Measures

In this chapter, we develop numerical summary measures for data sets consisting of a single variable.

When a data set contains more than one variable, the same numerical measures can be computed separately for each variable. However, in the two-variable case, we will also develop measures of the relationship between the variables.

If the measures are computed for data from a sample, they are called **sample statistics**.

If the measures are computed for data from a population, they are called **population parameters**.

A sample statistic is referred to as the **point estimator** of the corresponding population parameter.

Statistical software packages and spreadsheets can be used to develop the descriptive statistics presented in this chapter.

3.1 Mean

The most important measure of central location is the **mean**, the average of all the data values.

For a sample with *n* observations, the formula for the sample mean is as follows.

$$\bar{x} = \frac{\sum x_i}{n}$$ where x_i is the ith observation in a data set of size n

Example: consider the class size data for a sample of five college classes: 46, 54, 42, 46, 32.

Using the introduced notation, we have: $x_1=46$, $x_2=54$, $x_3=42$, $x_4=46$, and $x_5=32$.

To compute the sample mean, we write:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{\sum x_1 + x_2 + x_3 + x_4 + x_5}{n} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

The sample mean $\bar{x}$ is the point estimator of the **population mean**, μ.

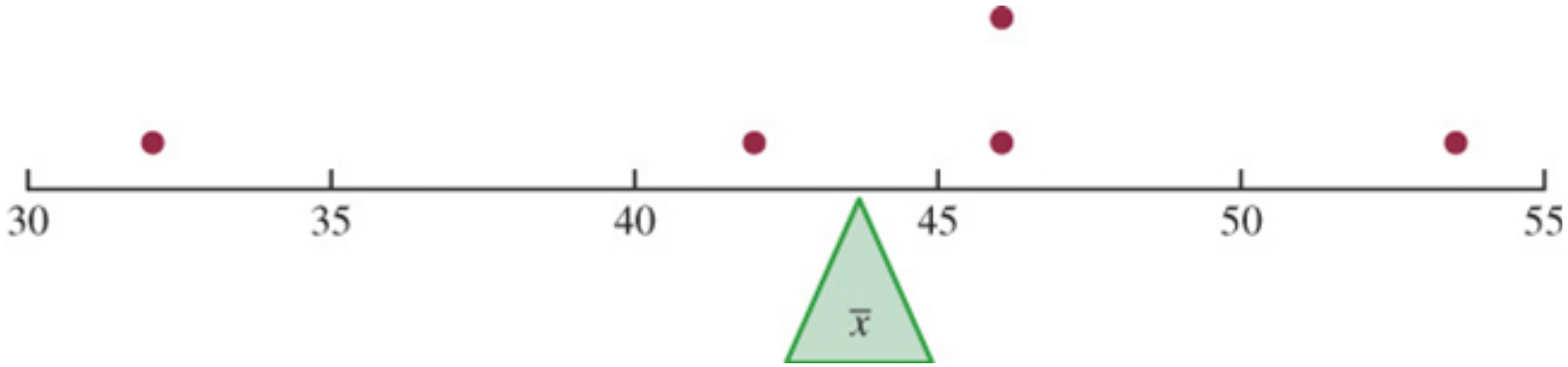
$$\mu = \frac{\sum x_i}{N}$$ where N is the size of the population

3.1 The Mean as the Center of Balance for the Dot Plot

If you view the horizontal axis of the dot plot as a long narrow board in which each of the dots has the same fixed weight, the mean is the fulcrum (or pivot point) that balances the dot plot.

This is how a see-saw on a playground works; the only difference is that the see-saw is pivoted in the middle so that as one end goes up, the other end goes down.

In the dot plot, we are locating the pivot point based on the location of the dots.



CENGAGE

3.1 Weighted Mean

In the formula for the mean, each data is given equal importance or weight.

In those instances where the mean is computed by giving each observation a weight that reflects its relative importance, we can calculate the **weighted mean** instead.

For a sample with *n* observations and weights w_i, the formula for the weighted mean is

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

Example: calculate the weighted mean for a sample of five purchases of raw materials listed to the right.

Purchase	Cost per Pound ($)	Weight (lbs)
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.95	1000
5	3.25	800

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{1{,}200(3.00) + 500(3.40) + 2{,}750(2.80) + 1{,}000(2.95) + 800(3.25)}{1{,}200 + 500 + 2{,}750 + 1{,}000 + 800} = \frac{18{,}500}{6{,}250} = \$2.96$$

CENGAGE

3.1 Median

The **median** is the value in the middle of a data set when data are arranged in ascending order.

To compute the median, first arrange the data in ascending order (smallest to largest value.)

a. If *n* is *odd*, the median is the middle value.

b. If *n* is *even*, the median is the average of the two middle values.

Example: consider the starting monthly salary for 12 business graduates.

First, we arrange the data in ascending order:

5710 5755 5850 5880 5880 5890 5920 5940 5950 6050 6130 6325

Because $n = 12$ is even, the median is the average of the 6th and 7th (middle) values.

$$Median = (5890 + 5920)/2 = 5905$$

Because the mean is influenced by extremely small and large data values, the median is the preferred measure of central location to report annual income and property value data.

3.1 Geometric Mean

The **geometric mean** is a measure of central location calculated by finding the nth root of the product of n values.

The general formula for the geometric mean, denoted $\bar{x}_g$, follows.

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\ldots(x_n)} = [(x_1)(x_2)\ldots(x_n)]^{1/n}$$

The geometric mean is often used in analyzing growth rates in financial data (where using the arithmetic mean will provide misleading results.)

It should be applied any time you want to determine the mean rate of change over several successive periods (be it years, quarters, weeks, etc.)

Other common applications include changes in populations of species, crop yields, pollution levels, and birth and death rates.

3.1 An Application of the Geometric Mean

Question: using 10 years of percentage annual returns compute how much \$100 invested in the fund at the beginning of year 1 would be worth at the end of year 10.

Year	Return (%)	Growth Factor
1	-22.1	0.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	-37.0	0.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

Because the percentage annual return for year 1 was −22.1%, the balance in the fund at the end of year 1 would be

$$\$100 - 22.1\%(\$100) = \$100(1 - 0.221) = \$100(0.779) = \$77.90$$

We refer to 0.779 as the **growth factor** for year 1.

We can generalize the result for year 1 to show the growth factor at the end of year 10.

$$(0.779)(1.287)(1.109)(1.049)(1.158)(1.055)(0.630)(1.265)(1.151)(1.021) = 1.334493$$

At the end of year 10, the initial investment would be worth $\$100(1.334493) = \$133.4493$

The mutual fund average growth is: $\bar{x}_g = \sqrt[10]{1.334493} = (1.334493)^{1/10} = 1.029275$

Thus, the fund average annual return is: $(1.029275 - 1)100\% = 2.9275\%$

CENGAGE

3.1 Mode

The **mode** of a data set is the value that occurs with the greatest frequency.

Example: consider again the starting monthly salary for 12 business graduates.

5710 5755 5850 5880 5880 5890 5920 5940 5950 6050 6130 6325

$5,880 is the mode because it is the only monthly starting salary that occurs more than once.

It may happen that the greatest frequency occurs at two or more different values.

In these instances, more than one mode exists.

In the presence of multiple modes, we define the following two cases:

- If the data have exactly two modes, the data are said to be *bimodal*.
- If the data have more than two modes, the data are said to be *multimodal*.

CENGAGE

3.1 Percentile

A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.

Admission test scores for colleges and universities are frequently reported in terms of percentiles.

The p**th percentile** of a data set is a value such that at least p percent of the items take on this value or less and at least $(100-p)$ percent of the items take on this value or more.

To calculate the pth percentile of a data set, we must first arrange the data in ascending order so that the smallest value in the data set is in position 1, the next one in position 2, and so on.

The **location** of the pth percentile, denoted L_p, is computed using the following equation:

$$L_p = \frac{p}{100}(n+1)$$

Now, we are ready to calculate the pth percentile. Let us do that with an example.

3.1 An Application of the Percentile

Compute the 80th percentile for the sample of 12 business graduates' starting salaries.

With the data arranged in ascending order, we indicate the position of each observation directly below its value.

	5710	5755	5850	5880	5880	5890	5920	5940	5950	6050	6130	6325
Position	1	2	3	4	5	6	7	8	9	10	11	12

The location of the 80th percentile is

$$L_{80} = \frac{p}{100}(n+1) = \frac{80}{100}(12+1) = 10.4$$

The interpretation of $L_{80} = 10.4$ is that the 80th percentile is 40% of the way between the values in position 10 and 11.

$$\text{80th percentile} = 6050 + 0.4(6130 - 6050) = 6050 + 0.4(80) = 6082$$

CENGAGE

3.1 Quartile

Quartiles are specific percentiles that divide the data set into four parts, with each part containing approximately 25% of the observations. Quartiles are defined as follows:

Q_1 = first quartile, or 25th percentile

Q_2 = second quartile, or 50th percentile (also the median)

Q_3 = third quartile, or 75th percentile

The procedure for computing percentiles can be also used to compute quartiles. Let us calculate Q_1 and Q_3 for the sample of 12 business graduates' starting salaries.

First, we calculate the locations

$$L_{25} = \frac{25}{100}(12 + 1) = 3.25 \quad \text{and} \quad L_{75} = \frac{75}{100}(12 + 1) = 9.75$$

The calculations of Q_1 and Q_3 follow:

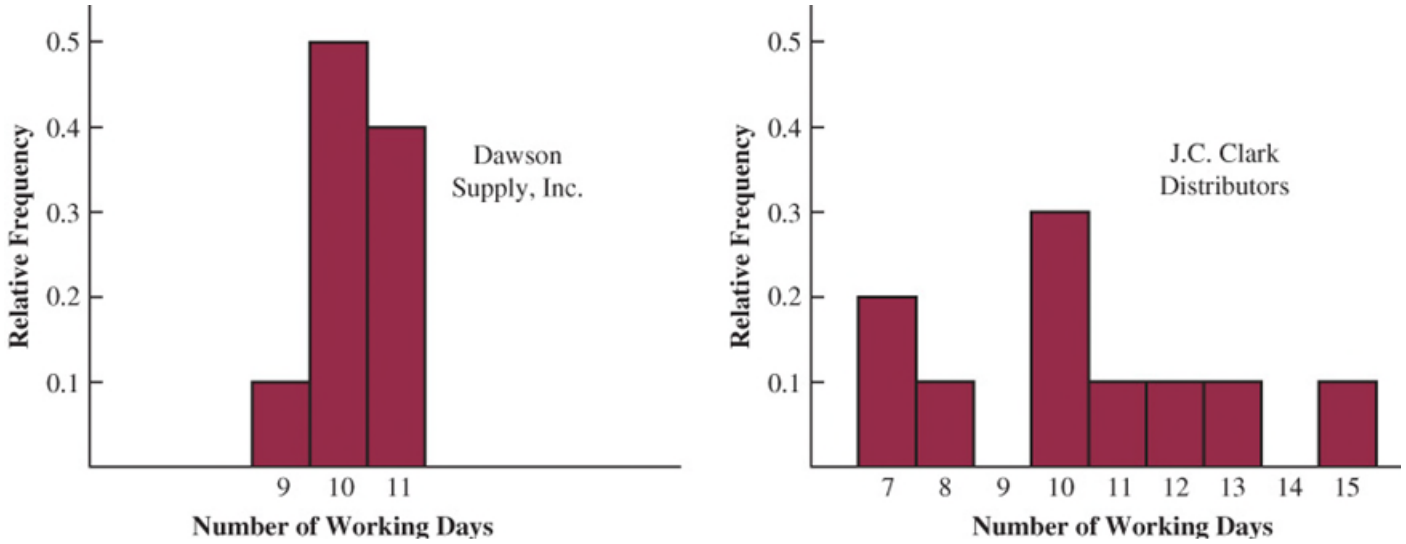
$$Q_1 = 5850 + 0.25(5880 - 5850) = 5857.5 \qquad Q_3 = 5950 + 0.75(6050 - 5950) = 6025$$

3.2 Measures of Variability

It is desirable to consider measures of variability (dispersion), as well as measures of location.

Consider the histograms for the number of days required to fill orders for two suppliers.

Although the mean number of days is 10 for both suppliers and the supplier to the right is able to fill some orders in as little as 7 to 8 days, the supplier to the left has a lower dispersion and demonstrates a higher degree of reliability in terms of making deliveries on schedule.



3.2 Range and Interquartile Range

Range

The **range** is the simplest measure of variability, and it is defined as

Range = Largest Value – Smallest Value

For the example of the business graduates' starting salaries, the range is

$$6325 - 5710 = 615$$

However, the range sensitivity to extreme data values makes it a poor choice to measure the dispersion in a data set.

Interquartile Range

The **interquartile range (IQR)** overcomes the dependency on extreme values considering the range for the middle 50% of the data.

The interquartile range is calculated as the difference between the third quartile, Q_3, and the first quartile, Q_1.

$$\boldsymbol{IQR = Q_3 - Q_1}$$

For the example of the business graduates' starting salaries, the *IQR* is

$$6025 - 5857.5 = 167.5$$

CENGAGE

3.2 Variance

The **variance** is a measure of variability that utilizes all the data.

The variance is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$ for a sample, μ for a population.)

The difference between each x_i and the mean is called a *deviation about the mean*.

In the computation of the variance, the deviations about the mean are squared.

Population Variance

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

Where *N* is the population size.

Sample Variance (*see notes)

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

Where *n* is the sample size.

CENGAGE

3.2 A Calculation of the Sample Variance

Let us calculate the sample variance of the class size for the sample of five college classes as presented in the previous section.

A summary of the data, including the computation of the deviations about the mean and the squared deviations about the mean, is shown in the table below.

The sum of squared deviations about the mean is $\sum(x_i - \bar{x})^2 = 256$.

With $n - 1 = 4$, the sample variance is

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

Note that, because the variance is a squared operator, its units are also squared.

In this case, the units are $(\text{students})^2$, and not students.

Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
46	44	4
54	44	100
42	44	4
46	44	4
32	44	144
		$\sum(x_i - \bar{x})^2 = 256$

3.2 Standard Deviation

The **standard deviation** is defined to be the positive square root of the variance.

Sample standard deviation: $\boldsymbol{s = \sqrt{s^2}}$

Population standard deviation: $\boldsymbol{\sigma = \sqrt{\sigma^2}}$

In the previous example, we calculated the sample variance of the class size for the sample of five college classes as $s^2 = 64$.

Thus, the sample standard deviation of the class size is

$s = \sqrt{64} = 8$ students

Because the standard deviation is the square root of the variance, the units of the variance, $(\text{students})^2$, are converted to students in the standard deviation.

Thus, the standard deviation is more easily compared to the mean and other statistics that are measured in the same units as the original data.

CENGAGE

3.2 Coefficient of Variation

The **coefficient of variation**, usually expressed as a percentage, measures how large the standard deviation is relative to the mean.

$$\left(\frac{\text{Standard Deviation}}{\text{Mean}} \times 100\right)\%$$

For the class size of the sample of five college classes, we found a sample mean of 44 and a sample standard deviation of 8.

The coefficient of variation is

$$\left(\frac{8}{44} \times 100\right)\% = 18.2\%$$

In words, the coefficient of variation tells us that the sample standard deviation is 18.2% of the value of the sample mean.

3.3 z-Scores

In addition to measures of location, variability, and shape, *measures of relative location* help us determine how far a particular value is from the mean.

The ***z*-score**, often called the *standardized value*, denotes the number of standard deviations, *s* a data value x_i is from the mean, $\bar{x}$.

$$z_i = \frac{x_i - \bar{x}}{s}$$

The *z*-scores for the class size data from the previous section are computed to the right.

For example, for $x_1 = 46$, the *z*-score is

$$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{46 - 44}{8} = \frac{2}{8} = 0.25$$

Number of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z–Score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2 / 8 = 0.25
54	10	10 / 8 = 1.25
42	2	–2 / 8 = -0.25
46	2	2 / 8 = 0.25
32	12	–12 / 8 = –1.50

3.3 Distribution Shape

Skewness is an important numerical measure of the shape of a distribution.

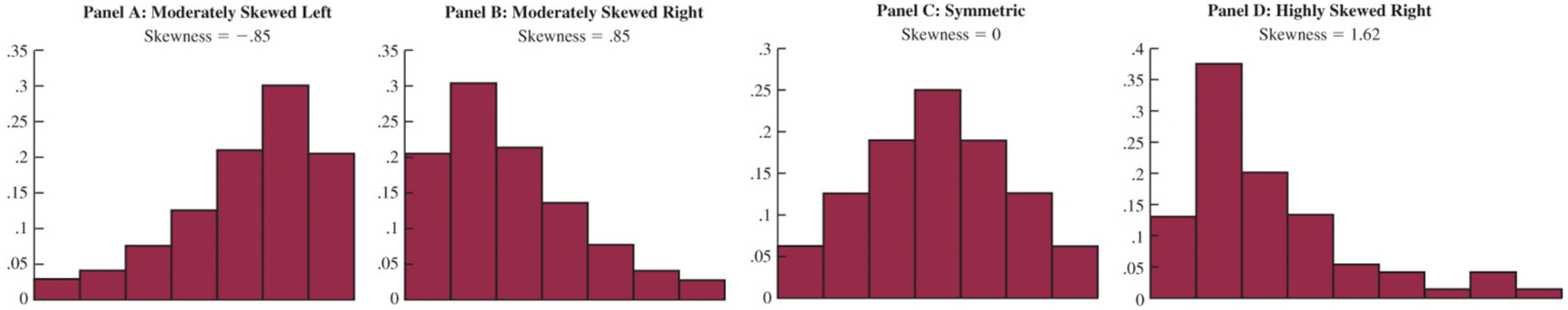
Because of its complexity, skewness is usually calculated with the help of statistical software (*see notes for the skewness formula.)

Panel A: a distribution moderately skewed to the left has negative skewness.

Panel B: a distribution moderately skewed to the right has positive skewness.

Panel C: a symmetric distribution has the mean equal to the median, and skewness = 0.

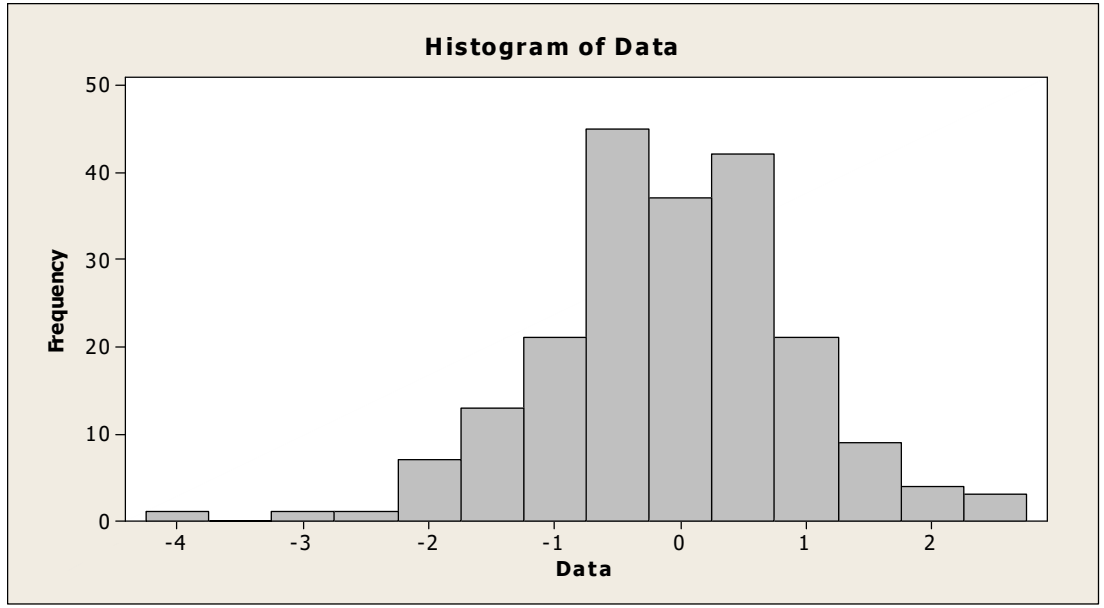
Panel D: a distribution highly skewed to the right has a larger positive skewness.

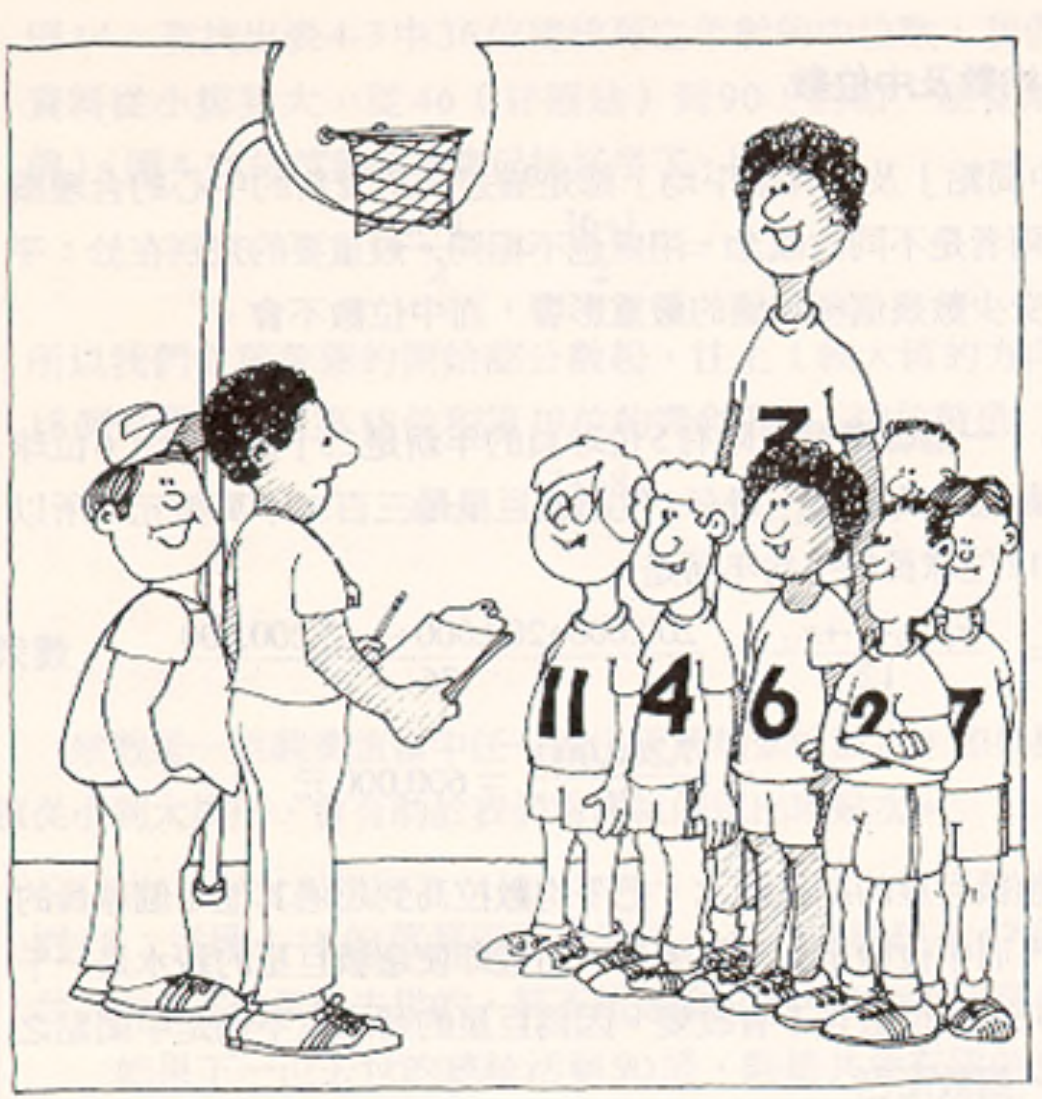


CENGAGE

平均數與中位數

- 左偏(Left-skewed)：當少數觀察值明顯小於一般觀察值，平均數將被這些觀察值拉下，但中位數較不受影響，此時中位數大於平均數。
- 少數觀察值較大時稱為右偏(Right-skewed)。





「我們是應該宣布我們的平均高度來嚇死對手，還是宣布我們的中位數高度來消除他們的戒心呢？」

3.3 Chebyshev's Theorem

Chebyshev's theorem enables us to make statements about the proportion of data values that must be within a specified number of standard deviations of the mean:

> **At least $(1 - 1/z^2)$ of the data values must be within *z* standard deviations of the mean, where *z* is any value greater than 1.**

Some of the implications of the theorem for $z = 2$ and 3 are:

- At least 75% of the data values must be within $z = 2$ standard deviations of the mean.
- At least 89% of the data values must be within $z = 3$ standard deviations of the mean.

Example: the midterm test scores for 100 students in a college business statistics course had $\bar{x} = 70$ and $s = 5$. How many students had test scores between 58 and 82?

We have: $z = (58 - 70)/5 = -2.4$, and $z = (82 - 70)/5 = +2.4$

Applying Chebyshev's Theorem with $z = 2.4$, we have: $1 - 1/z^2 = 1 - 1/2.4^2 = 0.826$

Thus, at least 82.6% of the scores are between 58 and 82.

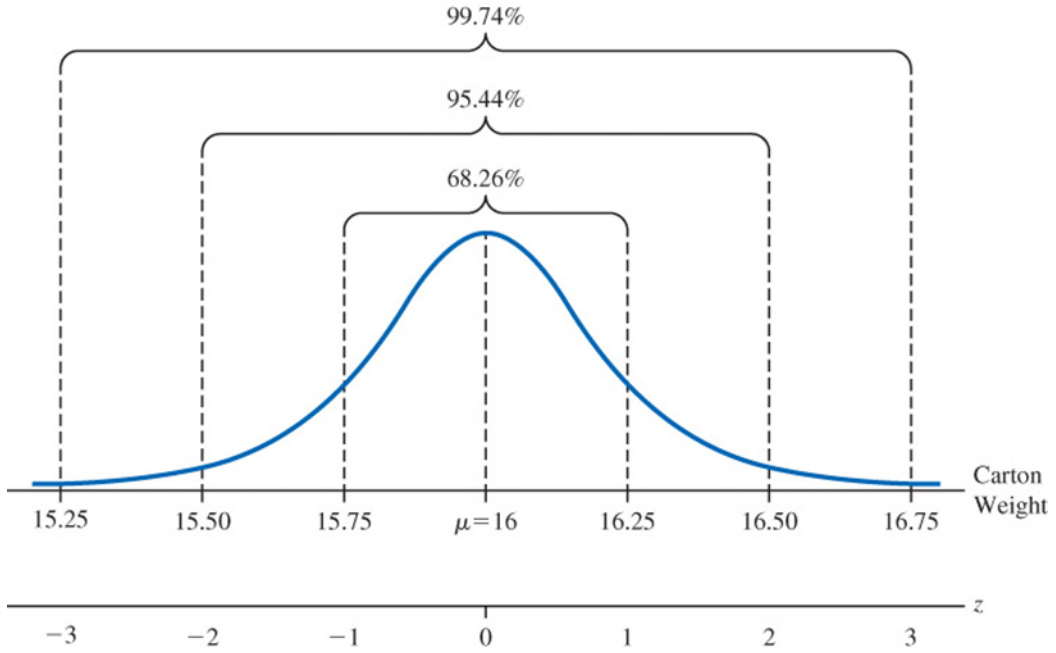
3.3 Empirical Rule

When the data are believed to approximate a symmetric bell-shaped distribution, the **empirical rule** can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.

For data having a bell-shaped distribution:

- ~68% of the data values will be within one standard deviation of the mean.
- ~95% of the data values will be within two standard deviations of the mean.
- Almost all of the data values will be within three standard deviations of the mean.

The empirical rule is based on the normal distribution, which we will cover in Chapter 6.



3.3 An Application of the Empirical Rule

Liquid detergent cartons are filled automatically on a production line. Filling weights have a bell-shaped distribution. If the mean filling weight is 16 ounces and the standard deviation is 0.25 ounces, we can use the empirical rule to conclude:

1. ~68% of the filled cartons will have weights between 15.75 and 16.25 oz.
2. ~95% of the filled cartons will have weights between 15.50 and 16.50 oz.

We can use this information to conclude approximately how many filled cartons will

- weigh between 16 and 16.25 oz:
  Because the distribution is symmetric, from point 1 we have $68\%/2 = 34\%$.
- weigh between 15.50 and 16 oz:
  Because the distribution is symmetric, from point 2 we have $95\%/2 = 47.5\%$.
- weigh less than 15.50 oz:
  Because 50% of the filled cartons have weights less than 16 oz. and 47.5% have weights between 15.50 and 16 ounces, it follows that 2.5% weigh less than 15.50 oz.

3.3 Detecting Outliers with *z*-Scores

An **outlier** is an unusually small or unusually large value in a data set.

Care should be taken when handling outliers, as they might be:

- an incorrectly recorded data value
- a data value that was incorrectly included in the data set
- a correctly recorded data value that belongs in the data set

A data value with a *z*-score less than –3 or greater than +3 might be considered an outlier.

For this example, refer to the starting monthly salary for the 12 business graduates, with mean, $\bar{x} = 5940$, and standard deviation, $s = 165.65$.

The smallest value in the data set, 5710, has $z = (5710 - 5940)/156.65 = -1.39$, and the largest value in the data set, 6325, has $z = (6325 - 5940)/156.65 = 2.32$.

Because all the values are within three standard deviations of the mean (*z*-scores within ±3), we conclude that there is no evidence of outliers.

3.3 Alternative Method for Detecting Outliers

Another approach to identifying outliers is based upon the values of the first and third quartiles (Q_1 and Q_3) and the interquartile range (*IQR*).

For this example, refer to the starting monthly salary for the 12 business graduates.

To use this method, we first compute the following lower and upper limits:

$$\text{Lower Limit} = Q_1 - 1.5(IQR) = 5857.5 - 1.5(167.5) = 5606.25$$

$$\text{Upper Limit} = Q_3 + 1.5(IQR) = 6025 + 1.5(167.5) = 6276.25$$

Looking at the starting monthly salary for the 12 business graduates, we see that:

There are no starting salaries lower than the Lower Limit = 5606.25.

There is one starting salary, 6325, that is greater than the Upper Limit = 6276.25.

Thus, using this alternate approach, 6325 is considered to be an outlier.

3.4 Five-Number Summary

In a **five-number summary**, five numbers are used to summarize the data:

1. Smallest value
2. First quartile (Q_1)
3. Median (Q_2)
4. Third quartile (Q_3)
5. Largest value

As an example, consider the monthly starting salary for the 12 business graduates.

The smallest value is 5710 and the largest value is 6325. We previously computed the quartiles ($Q_1 = 5857.5$; $Median = Q_2 = 5905$; and $Q_3 = 6025$).

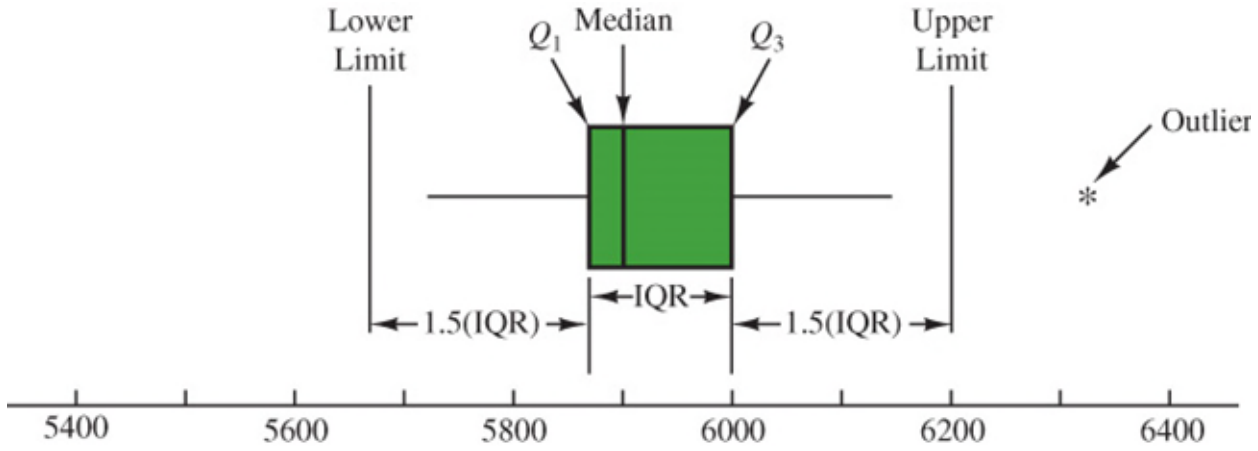
The five-number summary indicates that the starting salaries in the sample are between 5710 and 6325, approximately 50% of the starting salaries are between 5857.5 and 6025, and the median or middle value is 5905.

3.4 Boxplot

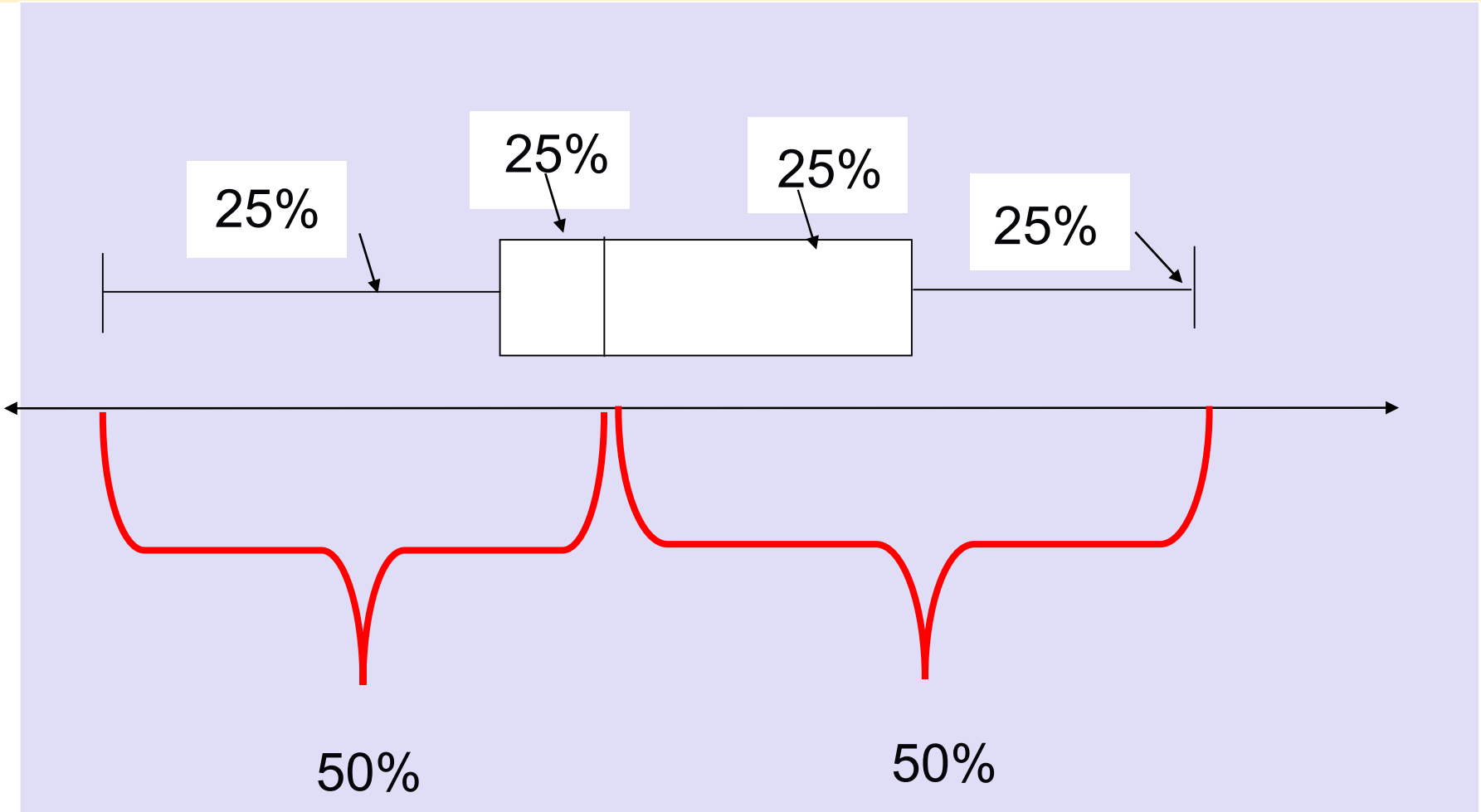
A **boxplot** is a graphical display of data based on a five-number summary.

The steps used to construct the boxplot for the monthly starting salary data are:

1. Draw a box between $Q_1 = 5857.5$ and $Q_3 = 6025$.
2. Draw a vertical line in the box at the location of the median, $Q_2 = 5905$.
3. Using $IQR = 167.5$, compute the lower and upper limits for detecting *outliers* (note that the limits are usually not drawn.)
4. Extend horizontal lines (*whiskers*) away from the box and up to the smallest and largest values that lie within the limits.
5. Use a small asterisk (*) to represent outliers.



Box-and-Whisker Plot (箱型圖)



- Boxplot也是常見的圖像呈現方式，只需要計算百分位數、四分位距，離群值為Q1–1.5*IQR及Q3＋1.5*IQR。

3.4 A Comparative Analysis Using Boxplots

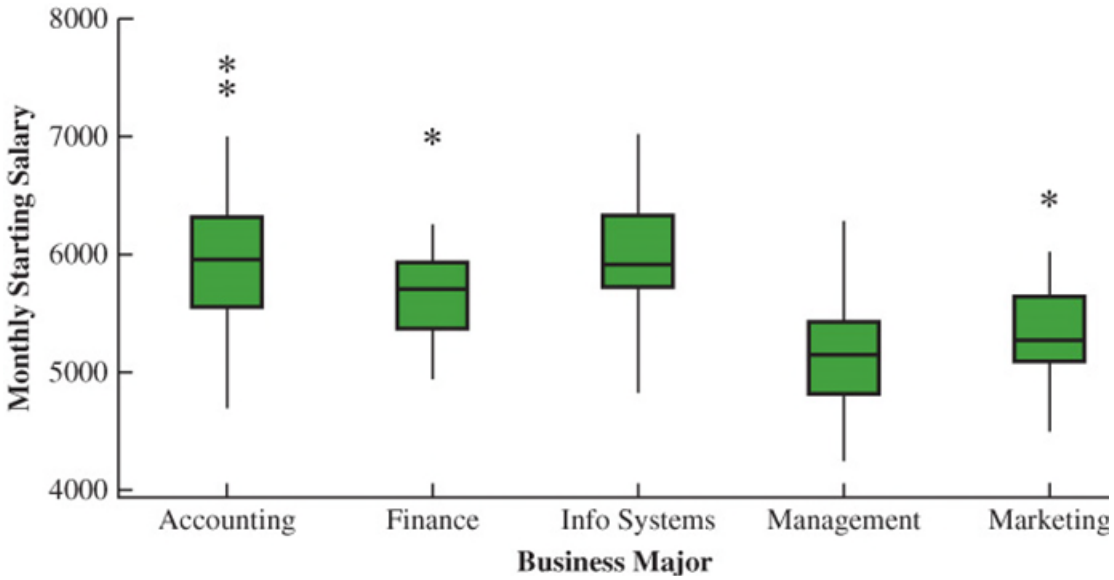
Boxplots can be used to provide a graphical summary of two or more groups and facilitate visual comparisons among the groups.

As an application, consider the major and starting salary data for a new sample of 111 recent business school graduates (DATAfile: *MajorSalaries*.)

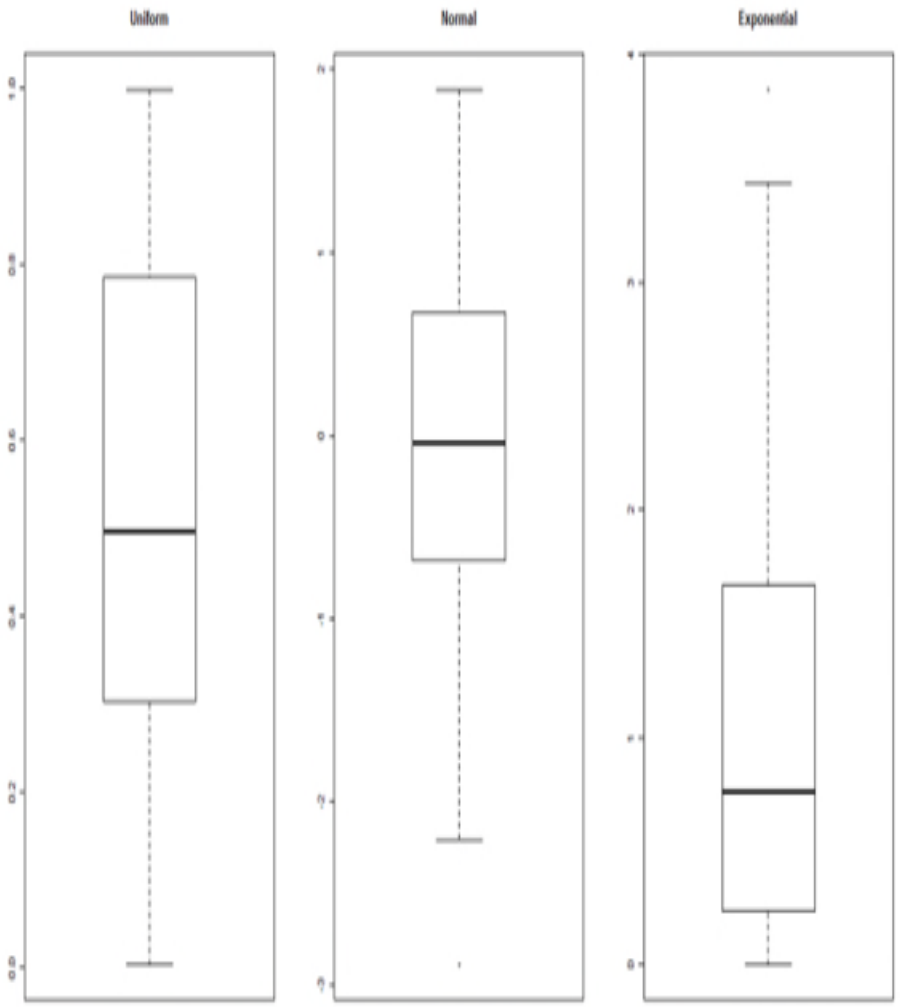
Note that the boxplots for each major have been drawn vertically, as is often the case.

A comparative analysis of the boxplots reveals:

- Accounting has the higher salaries, management and marketing the lower.
- Based on the median, Accounting and Information Systems have the higher median salaries, followed by Finance.
- High salary outliers exist for accounting, finance, and marketing majors.



基本統計量



以下是三種分配100個亂數的基本統計量：

(1)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.d.
.0037	.1967	.4577	.4601	.6920	.9707	.2821

(2)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.d.
-2.2060	-.4512	.1167	.1298	.9329	2.2570	.9507

(3)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.d.
.0214	.2483	.5749	.9590	1.2900	4.2170	.9856

註：第三筆資料明顯右偏；第一筆資料比第二筆資料更為「均勻」。

註：上述圖形為100個亂數的結果。

3.5 Measures of Association Between Two Variables

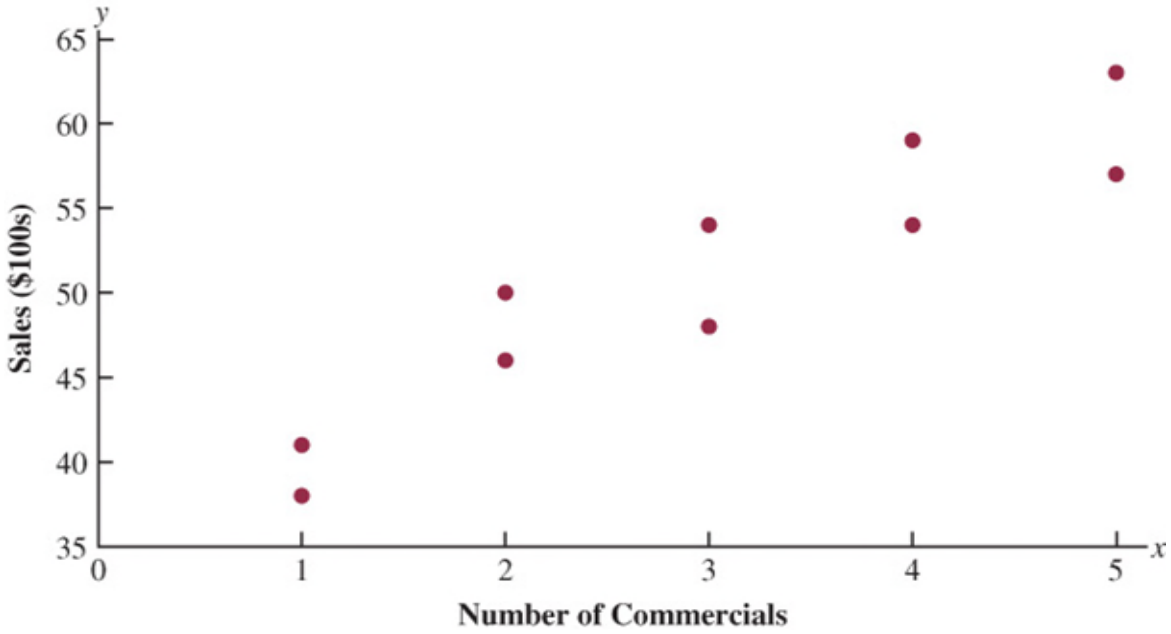
Often a manager or decision-maker is interested in the relationship between two variables.

Covariance and **correlation coefficient** are two descriptive measures of the *linear association* between two variables.

DATAfile: *Electronics*

As an application, we consider a sample of $n = 10$ weekly data for an electronics store in San Francisco.

The scatter diagram to the right suggests a *positive and linear* relationship between the number of weekend television commercials and the sales (in $100s) at the store during the following week.



3.5 Covariance

For a sample of size *n* with observations (x_i, y_i), and $i = 1 \ldots n$, the **sample covariance** is defined as

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

The table to the right shows the detailed calculations for the sample covariance of the electronics store data:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{9} = 11$$

The **population covariance** for a population of size N is

$$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
2	50	-1	-1	1
5	57	2	6	12
1	41	-2	-10	20
3	54	0	3	0
4	54	1	3	3
1	38	-2	-13	26
5	63	2	12	24
3	48	0	-3	0
4	59	1	8	8
2	46	-1	-5	5
30	510	0	0	99

CENGAGE

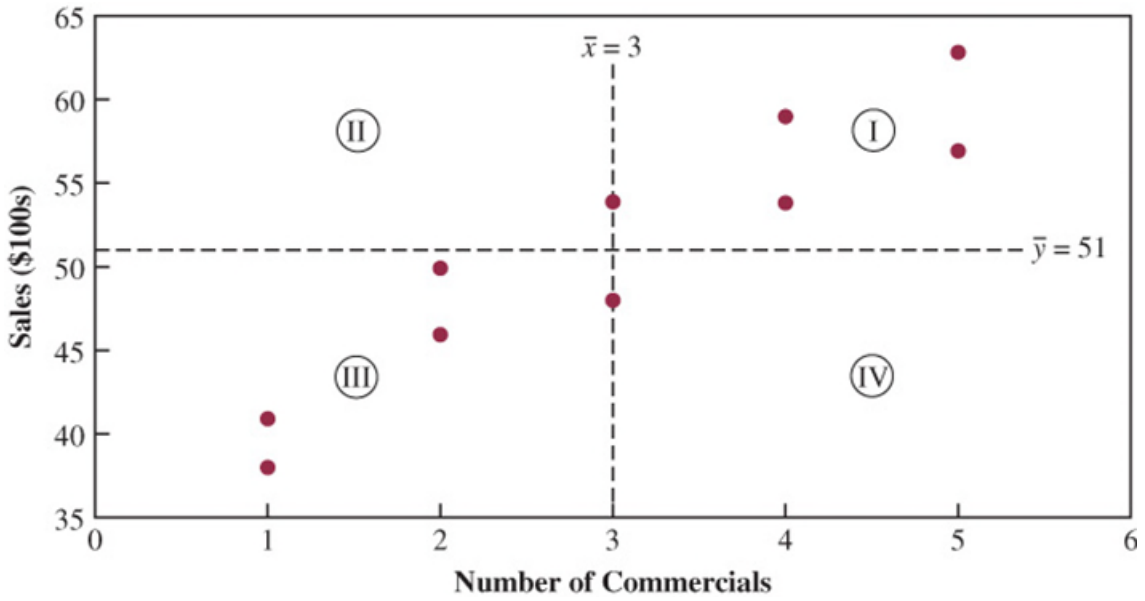
3.5 Interpretation of Sample Covariance

The figure shown to the right is the scatter diagram for the electronics store data, in which two dashed lines, a vertical line at $\bar{x} = 3$, and a horizontal line at $\bar{y} = 51$, dissect the diagram into four quadrants, labeled from I to IV.

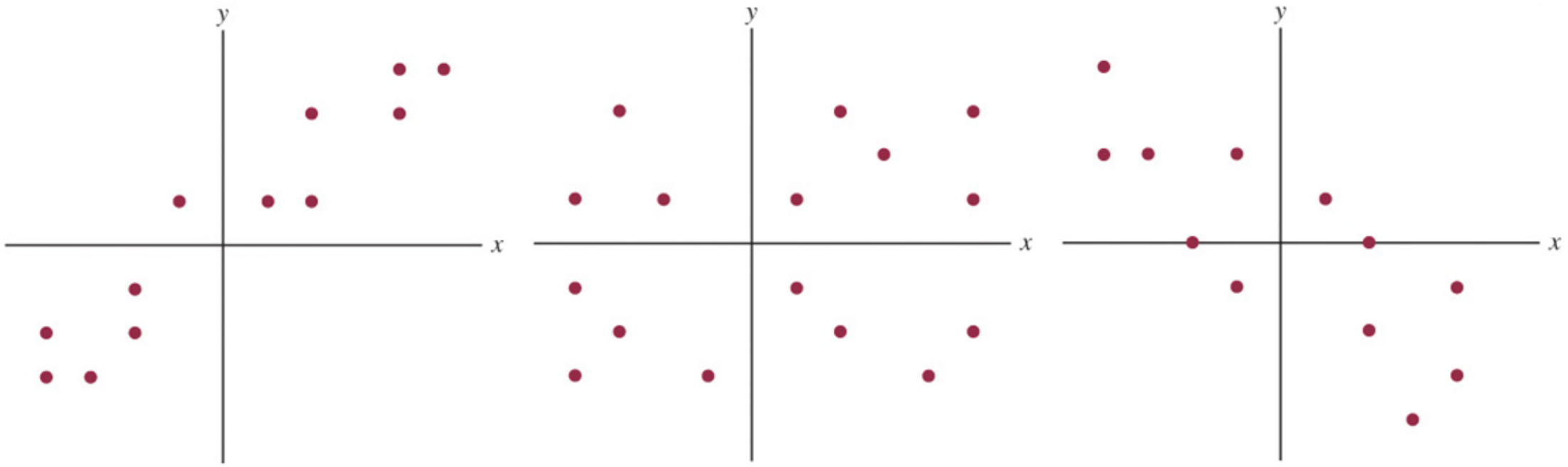
Because the covariance is related to the $(x_i - \bar{x})(y_i - \bar{y})$ term, it follows that quadrants:

- I and III have a *positive* contribution to the covariance, because $(x_i - \bar{x})$ and $(y_i - \bar{y})$ have the same sign.
- II and IV have a *negative* contribution to the covariance, because $(x_i - \bar{x})$ and $(y_i - \bar{y})$ have opposite sign.

The scatter diagram for the electronics store data shows that none of the data appear in quadrants II or IV. Thus, s_{xy} is positive.



3.5 Relationship Between Pattern and Covariance



s_{xy} Positive:
(x and y are positively linearly related)

s_{xy} Approximately 0:
(x and y are not linearly related)

s_{xy} Negative:
(x and y are negatively linearly related)

CENGAGE

3.5 Correlation Coefficient

The **correlation coefficient** is a measure of the relationship between *x* and *y* that is not affected by the units of measurement.

The *Pearson product moment correlation coefficient* is defined as

Sample Data

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Where

r_{xy} = sample correlation coefficient

s_{xy} = sample covariance

s_x = sample standard deviation of *x*

s_y = sample standard deviation of *y*

Population Data

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Where

ρ_{xy} = population correlation coefficient

σ_{xy} = population covariance

σ_x = population standard deviation for *x*

σ_y = population standard deviation for *y*

CENGAGE

3.5 Interpretation of the Correlation Coefficient

Let us now compute the sample correlation coefficient for the San Francisco electronics store.

$$s_x^2 = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1.49 \qquad s_y^2 = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7.93$$

We already know that $s_{xy} = 11$. Thus, the sample correlation coefficient is

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.49)(7.93)} = 0.93$$

- The correlation coefficient is always between −1 and +1.
- The sign of the correlation coefficient matches the direction of the association.
- The closer to +1, the stronger the positive linear association between *x* and *y*.
- The closer to −1, the stronger the negative linear association between *x* and *y*.
- The closer to 0, the weaker the linear association between *x* and *y*.

3.5 A Note on Linear Association

The correlation coefficient measures only the strength of the *linear association* between two quantitative variables.

The sample correlation coefficient for the data shown here is $r_{xy} = -0.007$ and indicates that there is no linear relationship between the two variables.

However, the scatter diagram provides strong visual evidence of a *nonlinear* relationship

CENGAGE

Summary

- In this chapter, we introduced descriptive statistics that can be used to summarize the location, variability, and shape of the distribution.
  - Numerical values obtained for a sample are called sample statistics.
  - Numerical values obtained for a population are called population parameters.
- We defined the mean, median, mode, weighted mean, geometric mean, percentiles, and quartiles as measures of location.
- Next, we presented the range, interquartile range, variance, standard deviation, and coefficient of variation as measures of variability or dispersion.
- We introduced skewness as a measure of the shape of the distribution, described how the mean and standard deviation are used to provide more information about data distribution (Chebyshev's theorem and the empirical rule), and to identify outliers.
- Finally, we introduced covariance and the correlation coefficient as measures of association between two variables.

CENGAGE

Case Problem 4: Heavenly Chocolates Website Transactions

Heavenly Chocolates manufactures and sells quality chocolate products at its plant and retail store located in Saratoga Springs, New York. Two years ago the company developed a website and began selling its products over the Internet. Website sales have exceeded the company's expectations, and management is now considering strategies to increase sales even further. To learn more about the website customers, a sample of 50 Heavenly Chocolate transactions was selected from the previous month's sales. Data showing the day of the week each transaction was made, the type of browser the customer used, the time spent on the website, the number of website pages viewed, and the amount spent by each of the 50 customers are contained in the file HeavenlyChocolates. A portion of the data are shown in Table 3.12.

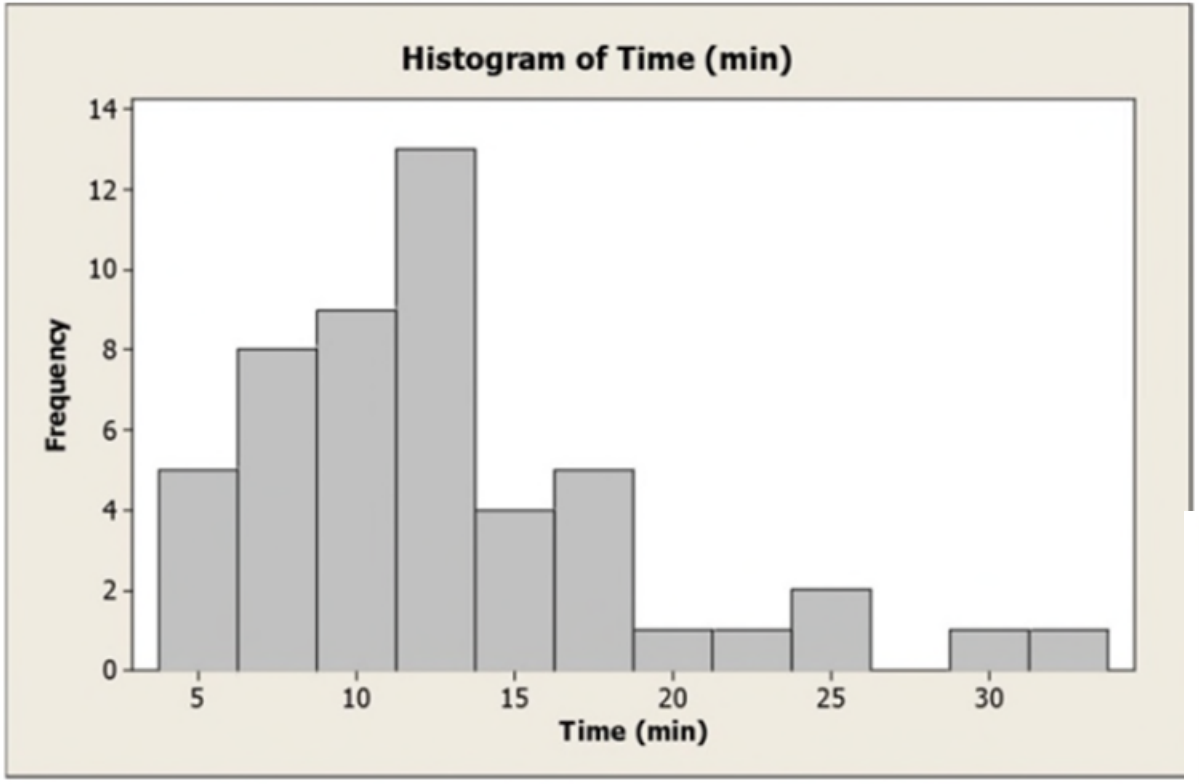
Heavenly Chocolates would like to use the sample data to determine if online shoppers who spend more time and view more pages also spend more money during their visit to the website. The company would also like to investigate the effect that the day of the week and the type of browser have on sales.

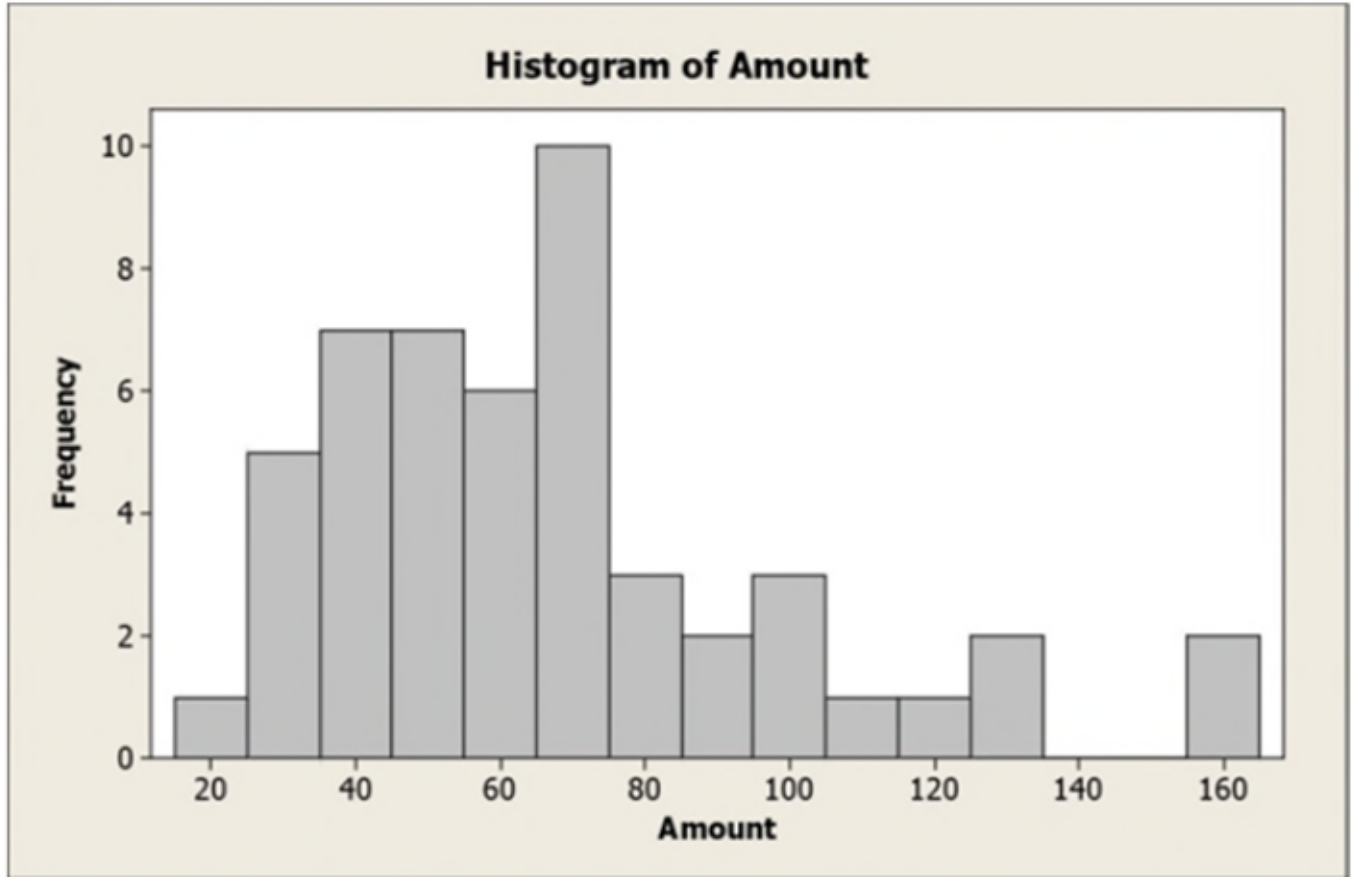
Table 3.12 A Sample of 50 Heavenly Chocolates Website Transactions

Customer	Day	Browser	Time (min)	Pages Viewed	Amount Spent ($)
1	Mon	Chrome	12.0	4	54.52
2	Wed	Other	19.5	6	94.90
3	Mon	Chrome	8.5	4	26.68
4	Tue	Firefox	11.4	2	44.73
5	Wed	Chrome	11.3	4	66.27
6	Sat	Firefox	10.5	6	67.80
7	Sun	Chrome	11.4	2	36.04
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
48	Fri	Chrome	9.7	5	103.15
49	Mon	Other	7.3	6	52.15
50	Fri	Chrome	13.4	3	98.75

	Time (min)	Pages Viewed	Amount Spent ($)
Mean	12.8	4.8	68.13
Median	11.4	4.5	62.15
Standard deviation	6.06	2.04	32.34
Skewness	1.45	.65	1.05
Range	28.6	8	140.67
Minimum	4.3	2	17.84
Maximum	32.9	10	158.51
Sum	640.5	241	3406.41

$$\text{Skewness} = \gamma_1 = \frac{E[(X - \mu)^3]}{\sigma^3}$$


Histogram of Time (min)
Frequency
0
2
4
6
8
10
12
14
5
10
15
20
25
30
Time (min)


Histogram of Amount
Frequency
0
2
4
6
8
10
20
40
60
80
100
120
140
160
Amount